Journal of Information Technology and Computer Science
Volume 11, Number 1, April 2026, pp. 77-90
Journal Homepage: www.jitecs.ub.ac.id

Comparing Audio and Visual Transfer Learning for Environmental Sound Classification

I Kadek Arya Augianta[1]
[1]Universitas Bali Internasional, Denpasar
[1]aryabisabikin@gmail.com
*Corresponding Author

Received 24 Desember 2026; accepted 16 May 2026

Abstract. Environmental Sound Classification (ESC) faces significant challenges related to data scarcity and the variability of unstructured acoustic signals. This study evaluates the effectiveness of a Visual Transfer Learning approach by converting audio signals into Mel-Spectrogram representations for classification using Computer Vision architectures. A comparative study was conducted on the ESC-50 dataset, comparing visual-based models (EfficientNet-B0, ResNet-50) with specialized audio models (Pre-trained Audio Neural Networks/PANNs). Experimental results show that EfficientNet-B0, optimized with MixUp augmentation, achieves the highest performance with an accuracy of 83.33% and an F1-Score of 83.50%. These results outperform ResNet-50 (80.00%) and significantly surpass the PANNs model (Cnn14), which only achieves 66.33%. The poor performance of PANNs indicates an issue of over-parameterization on small-scale datasets. Further validation using Gradient-weighted Class Activation Mapping (Grad-CAM) confirmed that the EfficientNet-B0 model is capable of precisely learning semantic features by distinguishing active sound patterns from silence and background noise. These findings confirm that lightweight visual architectures possess highly competitive transferability and generalization potential compared to massive audio models in the limited-data scenarios tested in this study.

Keywords: Environmental Sound Classification, EfficientNet-B0, Visual Transfer Learning, Grad-CAM, ESC-50.

1 Introduction

Sound is one of the largest information modalities in the human environment besides visual. Over the past decade, voice recognition technology has developed rapidly, transforming from mere speech recognition into a broader understanding of the acoustic environment, known as Environmental Sound Classification (ESC). This technology plays a vital role in the development of smart cities, automated security surveillance systems, biodiversity monitoring, and assistive tools for the hearing-impaired [1], [2].

Unlike human speech signals, which have a regular phoneme structure, environmental sounds are unstructured, vary in duration, and often overlap with background noise [3]. Standard datasets such as ESC-50, which contains 50 classes of environmental sounds, present their own challenges due to limited data (only 2,000 audio clips) and high intra-class variability, which can hinder the model's ability to capture the full complexity of environmental sounds [4], [5]. Conventional approaches that rely on hand-crafted features such as Zero Crossing Rate (ZCR) or simple statistical

features have proven inadequate for capturing the complexity of intricate environmental patterns [6], [7].

One of the main debates in this domain is the selection of feature extraction techniques. Mel Frequency Cepstral Coefficients (MFCC) have long been the gold standard in speech recognition. MFCC works by decorrelating the signal, which often results in the loss of spatial information that is important for deep learning models that rely on spatial features for accurate predictions [8]. On the other hand, Log-Mel Spectrogram offers a richer time-frequency visual representation, which allows the use of Convolutional Neural Networks (CNN) architecture to treat sound signals as digital images [9].

The next biggest challenge is the risk of overfitting on limited datasets. Several previous studies have offered solutions, but they have limitations. Mushtaq and Su proposed a CNN architecture built from scratch with conventional data augmentation [4]. Although it achieved good accuracy, this approach requires computationally intensive training of the model from scratch. On the other hand, the Transfer Learning strategy using the VGGish and YAMnet models has been applied in previous studies [10]. Although effective, VGG models tend to have very large and heavy parameters, making them less efficient for implementation in systems with limited resources [11]. There have been few studies exploring the balance between model efficiency, resistance to overfitting, and high curation using lighter modern architectures.

Therefore, the main novelty of this research lies in the integration of the EfficientNet-B0 architecture, which is known for its high parameter efficiency, with advanced data augmentation strategies, namely MixUp and SpecAugment. MixUp works by linearly mixing two audio signals and their labels, forcing the model to learn from inter-class uncertainty, an approach that has not been optimally applied in previous ESC-50 research. This strategy ensures data diversity and label fidelity by creating convex combinations of samples, effectively smoothing the decision boundaries. This innovation aims to maximize model generalization on limited data without overloading computational resources.

Finally, the main weakness of Deep Learning models is their black box nature. High accuracy without understanding what features the model has learned often casts doubt on the validity of the system [12]. Therefore, this study not only focuses on accuracy metrics, but also integrates interpretability analysis using Grad-CAM (Gradient-weighted Class Activation Mapping). This aims to visualize the spectral areas that are the focus of the model's attention, so that it can be ensured that the model performs classification based on relevant voice features, not based on noise or data artifacts [13].

The primary scientific contribution of this study lies in a critical analysis of the domain gap and parameter efficiency phenomena in cross-domain transfer learning. Unlike conventional comparative studies, this research investigates the extent to which the inductive bias of computer vision architectures pre-trained on ImageNet can generalize to non-object acoustic features. Our findings provide empirical evidence that for small-scale datasets such as ESC-50, the deployment of massive audio-native models (e.g., PANNs) tends to induce over-parameterization [14] that hinders convergence, a phenomenon closely linked to the generalization gap observed in high-capacity models when applied to sparse data environments [15]. Conversely, lightweight visual architectures offer superior regularization and higher feature extraction efficiency, representing a more balanced model parsimony. This study provides a robust methodological framework for developing environmental sound classification systems specifically tailored for resource-constrained edge devices, while

addressing the critical trade-off between model complexity and performance

2 Method

2.1 Dataset and Preprocessing

The study utilizes the ESC-50 dataset [16] as the primary benchmark. Chosen for its high class variability and inherent data scarcity, it serves as an ideal testbed for evaluating model generalization under constrained conditions. The dataset comprises 2,000 environmental audio clips (5 seconds each) strictly balanced into 50 semantic classes (exactly 40 samples per class) under five major macro-categories: animals, natural soundscapes, human non-speech sounds, interior/domestic sounds, and exterior/urban noises.

In the pre-processing stage, all audio clips were uniformly resampled to 32 kHz. This specific rate was selected based on prior studies [14], o offer an optimal trade-off between computational efficiency and spectral fidelity. According to the Nyquist-Shannon sampling theorem, this rate effectively captures frequencies up to 16 kHz, which encompasses the dominant acoustic energy of most environmental sounds while filtering out unnecessary ultra-high-frequency noise. Following resampling, strict linear amplitude normalization was applied to scale the temporal signals to the [-1, 1] range, ensuring input stability and preventing gradient explosion during the initial training phases [4].

To ensure robust and unbiased evaluation, the dataset was partitioned using stratified random sampling into three independent, non-overlapping subsets: Training (70%, 1,400 samples), Validation (15%, 300 samples), and Testing (15%, 300 samples). This stratification strategy maintains a perfectly balanced class distribution, allocating exactly 28 training, 6 validation, and 6 testing samples per class. Explicitly defining this constrained allocation relying on only 28 training samples per category is crucial, as it establishes the strict low-data regime that rigorously tests the models' susceptibility to over-parameterization and fundamentally justifies the necessity of the heuristic augmentation strategies applied in subsequent phases [17].

2.2 Feature Extraction and Spectral Representation

Given that raw audio signals inherently possess high dimensionality and large stochastic variability, transformation into the frequency domain is an imperative step to expose informative acoustic patterns [18]. The feature extraction process in this study initiates with the Short-Time Fourier Transform (STFT) to convert one-dimensional time-domain signals into two-dimensional time-frequency representations. Audio signals, resampled at 32 kHz, are segmented into short frames using a window size of 1024 samples (equivalent to 32 ms) with an inter-frame hop length of 320 samples (10 ms). To mitigate spectral leakage artifacts at inter-frame discontinuities, a Hann window function is systematically applied to each segment prior to the Fourier transform.

In determining the optimal spectral representation, this study evaluates Mel-Frequency Cepstral Coefficients (MFCC) as a baseline against the Log-Mel Spectrogram. Although MFCC utilizing the first 13 coefficients is highly efficient for concisely representing the spectral envelope in human speech, its reliance on the Discrete Cosine Transform (DCT) inherently decorrelates filterbank energies. Consequently, this process violently smooths out high-frequency harmonic details and fine-grained acoustic textures that contain crucial discriminative information for non-verbal environmental sounds. Therefore, the Log-Mel Spectrogram is established as the

primary input feature. By preserving the rich spectro-temporal structural integrity, it effectively bridges the domain gap, allowing acoustic data to be processed utilizing the robust spatial inductive biases of visual Convolutional Neural Networks (CNNs) [19], [20].

To generate this representation, the magnitude spectrum of the STFT is mapped onto the Mel scale utilizing a strictly defined filterbank of 64 Mel bins. This specific dimensionality is strategically chosen to seamlessly align the input tensor with the architectural requirements of the pre-trained PANNs (Cnn14) model. The Mel scale mathematically approximates the non-linear pitch resolution of human auditory perception, applying the following standard conversion formula:

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) \quad (1)$$

Where f represents the frequency in Hertz and m is the corresponding Mel frequency. To strictly enforce data quality prior to any augmentation or training phase, the frequency range is rigorously band-pass filtered. A lower bound of 50 Hz and an upper bound of 14,000 Hz are applied to eliminate sub-audible mechanical noise and irrelevant high-frequency artifacts, respectively. Finally, logarithmic scaling is applied to the spectrogram amplitude, converting the power units into a decibel (dB) scale. This operation not only mimics human logarithmic loudness perception but also ensures numerical stability, ultimately producing a standardized 2-dimensional feature matrix of size $(T \times 64)$ ready for network ingestion.

2.3 Data Augmentation Strategy

A fundamental challenge in training Deep Learning models on datasets with limited variability, such as ESC-50, is the risk of overfitting [21], where models tend to memorize specific noise patterns in the training data and fail to generalize to unseen data. To mitigate this risk and ensure strict methodological rigor, this study formulates a stochastic, on-the-fly data augmentation strategy meticulously designed to fulfill four critical criteria: data quality, diversity, privacy preservation, and label fidelity. This robust approach utilizes a synergistic combination of two techniques: SpecAugment and MixUp.

The first method, SpecAugment, is applied directly to the Log-Mel Spectrogram representation without altering the temporal duration of the audio. This technique functions as a regularization mechanism by randomly removing partial information through two masking schemes [22]. Frequency Masking, which covers a frequency band of width f taken from a uniform distribution $[0, F]$, and Time Masking, which covers a time segment of width t from a uniform distribution$[0, T]$. Procedurally, this mechanism guarantees high data diversity by simulating real-world acoustic occlusions and sensor variations, forcing the model to reconstruct incomplete representations rather than over-relying on dominant frequency peaks. Furthermore, operating purely within the spectral domain maintains strict data quality, as it preserves the fundamental harmonic structures of the environmental sounds without introducing the destructive, non-semantic noise artifacts commonly associated with aggressive raw waveform manipulations.

As a complementary strategy to refine the decision boundary between classes, this study implements the MixUp technique. Unlike conventional additive noise methods, MixUp works by synthesizing new samples through the convex combination of two randomly drawn audio samples (x_i, x_j) and their corresponding one-hot encoded labels (y_i, y_j) [23] . The mathematical formulation for this mixing process is defined as

follows:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \quad (2)$$

where λ is the mixing coefficient drawn from a Beta distribution, $\lambda \sim \text{Beta}(\alpha, \alpha)$. This protocol inherently ensures privacy preservation by strictly relying on internal data recombination; it maximizes the utility of the existing ESC-50 dataset without the risks associated with scraping external data or utilizing pre-trained generative models that may leak sensitive external audio. Simultaneously, the mathematical interpolation of the labels alongside the input signals guarantees absolute label fidelity. The synthetic samples remain perfectly aligned with their semantic boundaries, teaching the model to understand probabilistic transitions rather than rigid binary limits. This is particularly relevant for environmental sound classification, where acoustic superposition such as rain mixing with wind is a natural, highly faithful representation of the real world.

2.4 Model Design and Implementation

To systematically investigate the efficacy of cross-domain knowledge transfer, this study formulates an experimental design evaluating four distinct CNN architectures. These models were strategically selected to represent the dichotomy between training from scratch versus transfer learning across both Visual and Audio domains.

First, a Custom CNN serves as the lightweight baseline, trained entirely from scratch. It consists of four sequential convolutional blocks (Conv2D, BatchNorm, ReLU, MaxPooling) with a Dropout layer (0.3) applied before the fully connected classifier. This configuration is specifically designed to mitigate overfitting on the small-scale ESC-50 dataset while establishing a baseline for topological feature compatibility.

Second, Visual Transfer Learning is assessed using ImageNet-pretrained ResNet-50 and EfficientNet-B0 to examine the adaptability of visual inductive biases on acoustic spectrograms. While ResNet-50 relies on deep residual learning with skip connections to effectively prevent vanishing gradients in deep networks [24], EfficientNet-B0 employs a principled Compound Scaling mechanism [25]. This scaling method uniformly balances network depth, width, and resolution, providing a superior balance of representational power and model parsimony (requiring only ~5 million parameters compared to ResNet-50's ~25 million). Due to this architectural efficiency, EfficientNet-B0 was designated as the focal backbone for the subsequent multi-level augmentation ablation study (SpecAugment and MixUp).

Finally, PANNs (Cnn14) is implemented to represent the native Audio Domain [14]. Pre-trained on the massive AudioSet (comprising over 2 million clips), this VGG-based architecture is highly parameterized (containing ~81 million parameters). It is strategically included as a benchmark to determine whether native audio-learned features provide a definitive discriminative advantage, or if its massive capacity inherently induces over-parameterization when fine-tuned on the strictly constrained ESC-50 dataset.

2.5 Experiment Configuration and Performance Evaluation

The experiment was implemented using the PyTorch framework and the Librosa library on a CUDA-enabled GPU. The training process used the Adam optimizer [26] with an initial learning rate of $1 \times 10^{-4}$for a maximum of 50 epochs, and applied an

Early Stopping mechanism to prevent overfitting. Due to hardware limitations, a Gradient Accumulation strategy (physical batch size: 4, accumulation steps: 8) was applied to achieve an effective batch size of 32.

To ensure the reproducibility of the experiments, this study used a fixed random seed throughout the entire process of model weight initialization and data splitting. Each test scenario was run three times, and the results reported in this study represent the average of those iterations to ensure the stability of performance and the validity of the findings.

In the MixUp augmentation scheme, the standard categorical cross-entropy loss function is modified into a linear combination of the target labels to accommodate probabilistic mixing, as defined in Equation (3):

$$L_{mix} = \lambda \cdot L(p, y_i) + (1 - \lambda) \cdot L(p, y_j) \quad (3)$$

Where L denotes the standard Cross-Entropy function, p is the prediction distribution, y_i and y_j are the original labels, and λ is the mixing coefficient. Performance evaluation relies on Accuracy, Precision, Recall, and F1-Score. Given the balanced nature of ESC-50, the Macro-averaged F1-Score serves as the primary metric to ensure unbiased evaluation [27], calculated as:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (4)$$

Furthermore, qualitative validation employs Confusion Matrix to map error distributions and Grad-CAM [28]. to interpret model decisions ("Black Box"). Grad-CAM generates heatmaps highlighting crucial spectral regions based on the gradient of the target score (y^c) with respect to feature maps (A^k):

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (5)$$

3 Results and Discussion

3.1 The Influence of Feature Representation on Baseline Models

The initial stage of the experiment was dedicated to evaluating the crucial impact of feature representation selection on classification effectiveness. Using an identical Custom CNN architecture as a control variable, two fundamentally different types of acoustic features were tested: (1) Mel-Frequency Cepstral Coefficients (MFCC), which represent cepstral-based compressed features, and (2) Log-Mel Spectrogram, which preserves the holistic time-frequency structure. The main objective of this comparison is to validate which feature has the most optimal topological compatibility for processing by a convolutional neural network (CNN).

A summary of the quantitative performance of both features on the test data is presented in detail in Table 1.

Table 1. Comparison of Classification Performance between MFCC and Log-Mel Spectrogram Features in a Custom CNN Model.

Input Feature	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MFCC (*Baseline*)	56.33	63.56	56.33	55.97
Log-Mel Spectrogram	73.33	79.33	73.33	72.41

Based on empirical data in Table 1, a significant performance disparity was identified between the two approaches. The use of Log-Mel Spectrogram produced an

accuracy of 73.33%, recording an absolute advantage of +17% compared to the baseline MFCC, which remained stagnant at 56.33%.

The performance deficit in MFCC can be attributed to the intrinsic nature of the algorithm, which involves the Discrete Cosine Transform (DCT) stage. Although effective for speech data compression, the DCT process decorrelates the filterbank energy and eliminates fine-grained spectral details. However, it is these details that contain crucial discriminative information for distinguishing complex environmental sound classes. In contrast, the Log-Mel Spectrogram presents data in a 2D matrix format that is rich in texture and harmonic patterns. Given that the CNN architecture essentially functions as a visual feature extractor that relies on spatial correlation, the model is able to exploit the wealth of information in the spectrogram to distinguish sound classes with much higher precision. Based on these findings, the Log-Mel Spectrogram was established as the standard input for all subsequent experiments.

The superiority of Log-Mel representation is not only evident from the final score, but also from the dynamics of the learning process visualized in Figure 1. The graph compares the stability of validation during 20 training epochs.

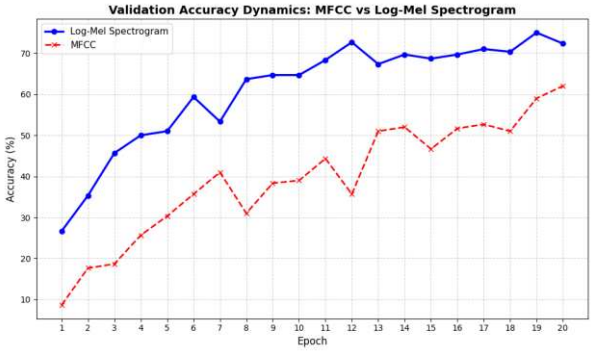

Figure 1. Validation Accuracy Dynamics: MFCC vs Log-Mel Spectrogram

The curve for the Log-Mel Spectrogram (blue line) shows a sharp and consistent upward trajectory since the early epochs, even exceeding the 50% accuracy threshold in just 4 epochs. This phenomenon indicates that the time-frequency topology structure in Log-Mel is highly compatible with the characteristics of local receptive fields in CNN convolution filters, enabling the model to learn efficiently.

In contrast, the MFCC curve (Red Dotted Line) shows high volatility, marked by a sawtooth pattern and a drastic performance collapse in the 7th and 17th epochs. This indication of gradient instability theoretically confirms that the loss of spatial correlation due to DCT compression in MFCC makes it difficult for the model to form a robust decision boundary. This graph provides empirical validation that the visual spectrogram-based approach (Log-Mel) is far more reliable than the cepstral approach (MFCC) in Deep Learning architecture.

3.2 Comparative Evaluation of Cross-Domain Architecture.

To address the performance limitations of the baseline model, the experiment was expanded by evaluating three comparative architectures from two different knowledge domains: ResNet-50 and EfficientNet-B0 (representing the visual/ImageNet domain of transfer learning), and PANNs Cnn14 (representing the audio/AudioSet domain). A comprehensive summary of the comparative performance

of all models is presented in Table 2.

Table 2. Comparative Evaluation of Cross-Architecture Classification Performance: Baseline Model, Audio-Based (PANNs), and Visual Transfer Learning.

Model	Domain Pre-training	Est. Parameter	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
PANNs (Cnn14)	Audio (AudioSet)	±81 Million	66.33	72.23	66.33	65.78
Custom CNN	*Scratch* (Log-Mel)	Lightweight	67.67	70.13	67.67	67.16
ResNet-50	Visual (ImageNet)	± 25 Million	80.00	82.55	80.00	79.53
EfficientNet-B0	Visual (ImageNet)	± 5 Million	82.00	84.43	82.00	82.11

The results in Table 2 show interesting performance dynamics. In the Visual Transfer Learning cluster, EfficientNet-B0 (82.00% accuracy) proved to outperform ResNet-50 (80.00%). This advantage is attributed to the Compound Scaling mechanism and the use of Swish activation in EfficientNet, which is more effective at capturing fine spectral features compared to standard deep residual architectures, while also being more efficient in parameter usage (±5 million vs. ±25 million).

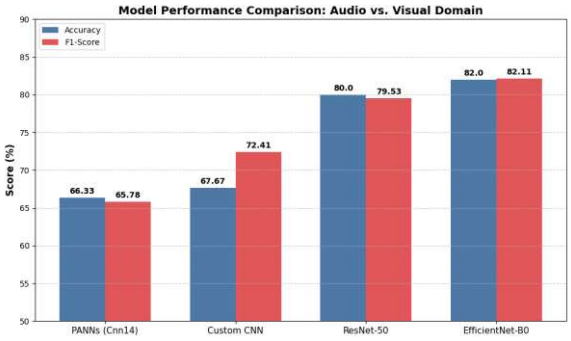

Figure 2. Comparison of Classification Performance (Accuracy and F1-Score) Across Model Architectures: A Comparative Evaluation between Audio Domain (PANNs) and Visual Domain (ResNet/EfficientNet) Approaches.

The visualization in Figure 2 highlights the significant performance disparity between the two knowledge domains, focusing on the Accuracy and F1-Score metrics to highlight the model's generalization capabilities. The bar chart demonstrates the dominance of visual transfer learning-based architectures (EfficientNet-B0 and ResNet-50), which consistently outperform PANNs models. The height of the bars for EfficientNet-B0 reflects the model's stability in maintaining performance balance, while the inferior position of PANNs highlights the ineffectiveness of using massive-capacity models on limited data. The suboptimal results achieved by PANNs (66.33%) in this experiment confirm the principle of Occam's Razor in deep learning, where increased model complexity does not necessarily translate to improved generalization, particularly in data-constrained environments. As suggested by Fan et al. [21], excessive model capacity without proportional data volume can lead the model to 'memorize' noise rather than learning robust semantic features, thereby widening the

generalization gap [27]. Furthermore, although the complete details are presented in Table 3, it should be noted that the low F1-score in PANNs (65.78%) is specifically triggered by a significant imbalance between Precision and Recall values. This indicates internal instability in handling class variability, in contrast to EfficientNet-B0, which achieves an optimal balance between both metrics by leveraging its more efficient inductive bias.

3.3 Ablation Analysis: Effectiveness of Augmentation Strategies

Following the establishment of EfficientNet-B0 as the most optimal backbone architecture, the focus of the experiment shifted to testing data regularization strategies to further push the limits of model generalization. An ablation study was conducted in stages by integrating two spectral augmentation techniques: SpecAugment, which works by randomly masking frequency and time ranges to force the model to learn stronger features, and MixUp, a linear interpolation technique that combines two audio samples and their labels according to the equation $x' = \lambda x_i + (1 - \lambda)x_j$ to smooth the decision boundary. The incremental performance improvements resulting from the integration of these strategies are summarized in Table 3.

Table 3. Impact of Multi-Level Data Augmentation on the Performance of the EfficientNet-B0 Model.

Model Configuration	Augmentation Method	Accuracy (%)	F1-Score (%)	Δ Accuracy
EfficientNet-B0 (*Baseline*)	*None*	82.00	82.11	–
EfficientNet-B0	+ SpecAugment	82.67	82.85	+0.67%
EfficientNet-B0 (*Final*)	+ SpecAugment + MixUp	83.33	83.50	+1.33%

The empirical data in Table 3 validates that data augmentation interventions have a consistent positive impact on model performance trajectories. The implementation of SpecAugment provides an initial increase of +0.67%, boosting accuracy to 82.67%. This improvement indicates that the masking mechanism, or partial concealment of spectral information, effectively prevents the model from relying too heavily on dominant frequency features alone, forcing it to reconstruct and recognize the spectral context more holistically. Peak performance was then achieved when the MixUp technique was integrated, which escalated the final accuracy to 83.33% with an F1-Score of 83.50%.

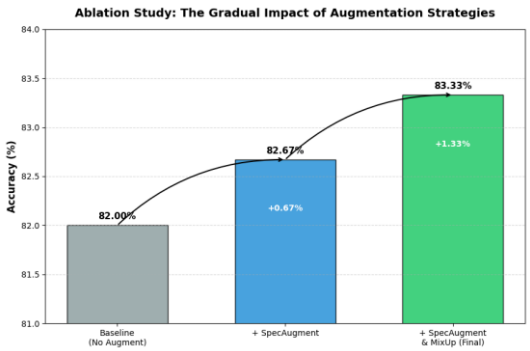

Figure 3. Ablation Study: The Gradual Impact of Augmentation Strategies

MixUp proved crucial in mitigating the risk of over-confidence in the model by

introducing synthetic mixed samples, which mathematically forced the model to build more linear and smooth inter-class prediction transitions. Thus, the combination of EfficientNet-B0 with hybrid augmentation (SpecAugment and MixUp) was established as the final configuration that was not only accurate but also highly robust against data variability.

The dynamics of this performance improvement are further visualized in Figure 3, which presents a summary of the ablation study in a step-wise improvement format. This visualization highlights the marginal but significant contribution of each stage of intervention. The baseline phase (visualized with gray bars) shows a solid starting point at 82.00%, which then escalates in the spectral regularization phase (blue bars) due to the application of SpecAugment. The graph peaks in the final generalization phase (green bars) with MixUp integration, resulting in a total increase of +1.33% from the initial condition. Although numerically moderate, this increase is statistically significant given the difficulty of increasing accuracy when the model has reached saturation above the 80% threshold. This provides empirical evidence that the applied signal manipulation strategy successfully strengthens the model's decision boundaries, making it the most robust solution in this series of studies.

3.4 Strategic Discussion on Data Augmentation: Heuristic vs. Generative

Following the recent paradigm shift in data augmentation strategies, as comprehensively reviewed by Cui et al. [29], this study acknowledges the burgeoning role of Generative AI such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) in synthesizing training samples to alleviate data scarcity. However, a critical synthesis of our methodological choices reveals that for Environmental Sound Classification (ESC) within a low-data regime like the ESC-50 dataset, heuristic-based spectral augmentations provide superior methodological rigor and feature fidelity compared to purely generative approaches.

The primary challenge with implementing Generative AI in small-scale acoustic datasets is the high risk of mode collapse and the potential introduction of "hallucinated" artifacts that do not align with true physical sound propagation. High-fidelity synthesis requires substantial baseline data to accurately learn the underlying data distribution [30]. When trained on sparse datasets, generative models often struggle to maintain the semantic integrity of the acoustic signals, producing non-semantic noise that can degrade the classifier's generalization capabilities.

In contrast, the heuristic strategies employed in this study directly address these vulnerabilities. The implementation of SpecAugment operates through strategic structural masking of the time-frequency manifold. This mechanism directly simulates acoustic occlusions found in real-world environments without altering the fundamental harmonic structure of the signal [31]. Furthermore, the integration of MixUp serves as a mathematically grounded form of manifold smoothing. By creating convex combinations of audio samples and their corresponding labels, MixUp forces the neural network to learn linear transitions between classes, effectively preventing the over-confidence commonly observed in over-parameterized models [32].

This mathematically controlled transition ensures high label fidelity, as the augmented samples remain strictly anchored to the original acoustic features rather than being stochastically generated. Ultimately, this controlled heuristic strategy maximizes information gain while preserving data integrity. This methodological rigor directly facilitates the high level of interpretability demonstrated in our Grad-CAM analysis in

the subsequent section, proving that the model identifies genuine semantic patterns rather than idiosyncratic generative artifacts.

3.5 Error Analysis and Model Interpretability

As a final validation step to ensure the integrity of the decisions made by the model, an in-depth evaluation was conducted on EfficientNet-B0, which had been trained using the MixUp Augmentation strategy. This analysis began with an examination of the error distribution using a Confusion Matrix on the test set, as visualized in Figure 4.

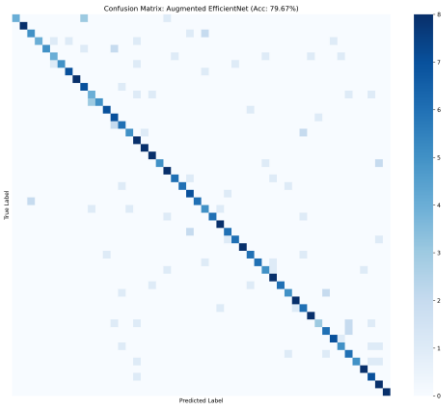

Figure 4. Confusion Matrix showing random error patterns

The visual pattern formed on the matrix shows a strong density concentration on the main diagonal line, reflecting a high level of correct predictions (True Positives). Conversely, classification errors (misclassification) represented by cells outside the main diagonal appear sporadically scattered and do not form a systematic confusion cluster pattern in certain class pairs. This phenomenon of random error distribution indicates that the model has excellent generalization capabilities and does not suffer from specific bias towards certain sound categories, a crucial achievement given the high complexity and similarity of features between acoustic classes in the ESC-50 dataset.

Next, internal validity was audited using Gradient-weighted Class Activation Mapping (Grad-CAM) visualization to ensure that predictions were based on relevant semantic features. Figure 5 presents a comparative analysis between correct and incorrect predictions. In the case of correct predictions (top row, class 'brushing_teeth'), the heatmap shows the model's attention highly localized to the active signal area. In fact, in samples with short signal durations (Figure 5 top-right), the model intelligently suppresses attention weights in areas of silence, proving that the model successfully isolates sound objects from the background.

Conversely, analysis of misclassification cases (bottom row) reveals the root cause of ambiguity. In the 'door_wood_knock' sample predicted as 'fireworks' (Figure 6 bottom left), Grad-CAM shows that the model's attention is spread across the entire impulsive transient pattern, which does indeed have spectral characteristics similar to the sound of fireworks. This confirms that the model's error is not caused by a failure to capture features, but rather by the inherent feature overlap between classes in the challenging ESC-50 dataset.

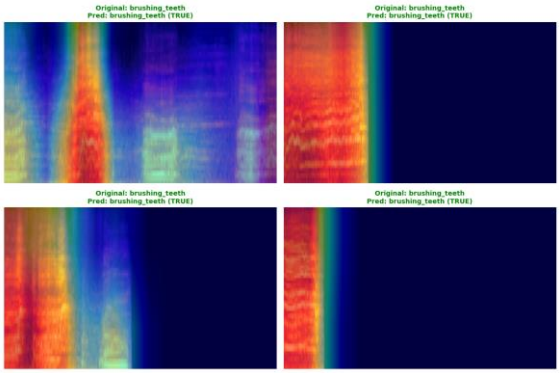

Figure 5. Grad-CAM visualization showing the model's focus on active sound segments (red/yellow) while successfully ignoring background silence (blue), confirming semantic learning capability

4 Conclusion

This study concludes that the Visual Transfer Learning approach using EfficientNet-B0 is a highly competitive solution for environmental sound classification with limited data, achieving the highest accuracy of 83.33% among the models evaluated in this study. These results surpass the ResNet-50 architecture (80.00%) and significantly outperform PANNs audio models (66.33%), which were found to suffer from over-parameterization when trained on small-scale datasets. Validation via Grad-CAM confirms that the model is capable of learning semantic features with precision by distinguishing active sound signals from background noise, which simultaneously demonstrates the reliability of the spectrogram transformation into the visual domain.

Although promising results were achieved, this study acknowledges the limitations of using the single ESC-50 dataset. As a next step, validation on broader and more diverse datasets, such as UrbanSound8K or AudioSet, is highly recommended to test the model's robustness and generalization capabilities on a larger scale. Additionally, future research could explore model compression techniques (pruning or quantization) to enable the EfficientNet-B0 architecture to be efficiently implemented on edge devices (IoT). Finally, testing the model's resilience in dynamic and noisy acoustic environments remains a key priority for validating the system's performance in the real world.

References

1. K. G. Dharmasiri, C. V Rathnasooriya, M. K. Balasuriya, L. N. Yapa, D. I. De Silva, and T. Thilakarathne, "Enhancing Environmental Awareness for Hard of Hearing Individuals: A Mobile Application Approach," in *Lecture Notes in Networks and Systems*, 2025, vol. 1444 LNNS, pp. 307–317. doi: 10.1007/978-981-96-6932-5_23.
2. R. Souadek and N. Guellil, "Environmental sound classification based on STFT features and Convolutional Neural Networks (CNNs)," *Prz. Elektrotechniczny*, vol. 101, no. 7, pp. 158–165, 2025, doi: 10.15199/48.2025.07.25.
3. M. Bhavya and M. R. Anala, "Deep Learning Approach for Sound Signal Processing," 2022. doi: 10.1109/INCOFT55651.2022.10094337.
4. Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Appl. Acoust.*, vol. 167, 2020, doi: 10.1016/j.apacoust.2020.107389.
5. M. M. Nasef, M. M. Nabil, and A. M. Sauber, "Multiclass environmental sound classification model based on adding residual connections to self-attention layers,"

Multimed. Tools Appl., vol. 83, no. 28, pp. 71359–71377, 2024, doi: 10.1007/s11042-024-18421-7.
6. W. Jin, X. Wang, and Y. Zhan, "Environmental Sound Classification Algorithm Based on Region Joint Signal Analysis Feature and Boosting Ensemble Learning," *Electronics*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223743.
7. B. Paul *et al.*, "Spoken word recognition using a novel speech boundary segment of voiceless articulatory consonants," *Int. J. Inf. Technol.*, vol. 16, no. 4, pp. 2661–2673, 2024, doi: 10.1007/s41870-024-01776-3.
8. R. Akter, M. R. Islam, S. K. Debnath, P. K. Sarker, and M. K. Uddin, "A hybrid CNN-LSTM model for environmental sound classification: Leveraging feature engineering and transfer learning," *Digit. Signal Process. A Rev. J.*, vol. 163, 2025, doi: 10.1016/j.dsp.2025.105234.
9. D. P. Goel, K. Mahajan, N. D. Nguyen, N. Srinivasan, and C. P. Lim, "Towards an efficient backbone for preserving features in speech emotion recognition: deep-shallow convolution with recurrent neural network," *Neural Comput. Appl.*, vol. 35, no. 3, pp. 2457–2469, 2023, doi: 10.1007/s00521-022-07723-2.
10. T. H. Rochadiani, Y. Arifin, D. Suhartono, and W. Budiharto, "Exploring Transfer Learning Approach for Environmental Sound Classification: A Comparative Analysis," in *2024 International Conference on Smart Computing, IoT and Machine Learning, SIML 2024*, 2024, pp. 112–116. doi: 10.1109/SIML61815.2024.10578248.
11. N. Zakaria and Y. M. M. Hassim, "Improved Image Classification Task Using Enhanced Visual Geometry Group of Convolution Neural Networks," *Int. J. Informatics Vis.*, vol. 7, no. 4, pp. 2498–2505, 2023, doi: 10.30630/joiv.7.4.1752.
12. V. Buhrmester, D. Münch, and M. Arens, "Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 4, pp. 966–989, 2021, doi: 10.3390/make3040048.
13. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.
14. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, pp. 2880–2894, Nov. 2020, doi: 10.1109/TASLP.2020.3030497.
15. B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, "The role of over-parametrization in generalization of neural networks," 2019. [Online]. Available: https://openreview.net/forum?id=BygfghAcYX
16. K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 1015–1018. doi: 10.1145/2733373.2806390.
17. H. Purwins, B. Li, T. Virtanen, J. Schl\&\#x00FC;ter, S. Chang, and T. N. Sainath, "Deep Learning for Audio Signal Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 13, pp. 206–219, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:132199224
18. S. Dalmia and M. Rege, "Anomalous Sound Pattern Detection for Machine Health Monitoring," in *Communications in Computer and Information Science*, 2024, vol. 2127 CCIS, pp. 44–60. doi: 10.1007/978-3-031-68617-7_4.
19. Y. Cai and W. Xu, "The best input feature when using convolutional neural network for cough recognition," in *Journal of Physics: Conference Series*, 2021, vol. 1865, no. 4. doi: 10.1088/1742-6596/1865/4/042111.
20. M. Wang *et al.*, "Environmental Sound Recognition Based on Double-input Convolutional Neural Network Model," in *Proceedings of 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology, ICCASIT 2020*, 2020, pp. 620–624. doi: 10.1109/ICCASIT50869.2020.9368517.
21. Y. Fan, H. Huang, and H. Han, "Quantifying overfitting in deep learning," *Anal. Appl.*, vol. 23, no. 5, pp. 705–729, 2025, doi: 10.1142/S0219530525400068.

22. X. Zhang, H. Xing, M. Song, D. Takeuchi, N. Harada, and S. Makino, "Prediction-error-based Adaptive SpecAugment for Fine-tuning the Masked Model on Audio Classification Tasks," 2024. doi: 10.1109/APSIPAASC63619.2025.10848691.
23. L. Cances, E. Labbé, and T. Pellegrini, "Comparison of semi-supervised deep learning algorithms for audio classification," *Eurasip J. Audio, Speech, Music Process.*, vol. 2022, no. 1, 2022, doi: 10.1186/s13636-022-00255-6.
24. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-December, pp. 770–778. doi: 10.1109/CVPR.2016.90.
25. M. Tan and Q. V Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICML 2019*, 2019, vol. 2019-June, pp. 10691–10700. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077515832&partnerID=40&md5=b8640eb4e9a606d0067b4a420ca73df1
26. R. Liu, T. Wu, and B. Mozafari, "Adam with bandit sampling for deep learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 2020-December. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108440289&partnerID=40&md5=ead956545ae74e1ef924db41eeb17b8b
27. R. Diallo, C. Edalo, and O. O. Awe, "Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score," in *STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics and Health*, vol. Part F4005, 2025, pp. 283–312. doi: 10.1007/978-3-031-72215-8_12.
28. R. Mothkur, P. Soubhagyalakshmi, and C. B. Swetha, "Grad-CAM Based Visualization for Interpretable Lung Cancer Categorization Using Deep CNN Models," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 3, pp. 567–580, 2025, doi: 10.35882/jeeemi.v7i3.690.
29. L. Cui, H. Li, K. Chen, L. Shou, and G. Chen, *Tabular Data Augmentation for Machine Learning: Progress and Prospects of Embracing Generative AI*. 2024. doi: 10.48550/arXiv.2407.21523.
30. J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3313–3332, 2023, doi: 10.1109/TKDE.2021.3130191.
31. C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.
32. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083953830&partnerID=40&md5=eddbbd01a1c41e4fcf39458cd829e2a9